

R: Introduction to Basic Features

Helen Lockstone
hel23@well.ox.ac.uk



iT Centre
Learning

iT
services



Course Overview



R is a powerful software environment for all kinds of statistical data analysis, and has become a popular and widely used tool in many disciplines. To harness its capabilities, an R user needs to be able to do all of the following:

1. Program instructions in the R language
2. Understand R's data structures, functions and behaviour
3. Use appropriate statistical tests and models for their data
4. Interpret the output correctly

On this course we focus on the first 2 items, introducing the R programming language and starting to handle data in the R environment. The course is intended to help those new to programming in general get acquainted with using R by helping overcome some of the barriers created by technical jargon and notation in user manuals, help pages and error messages.

About the tutor



- Lead the Bioinformatics Core at Wellcome Centre for Human Genetics and have many years experience using R for data analysis, particularly gene expression data.
- I and other colleagues have developed and taught a range of R and Bioinformatics Data Analysis courses to DPhil students and post-doctoral researchers over the past few years.

Schedule



09:30 - 10:00	Introducing the R environment, basic commands and data types
10:00 - 10:30	Understanding variables, functions and arguments
10:30 - 10:40	Tea/coffee break
10.40 - 11:30	Setting the working directory, reading in and accessing data
11.30 - 12:30	Performing simple data analysis and plotting

Course format: informal workshop-style - please feel free to ask questions at any time.

Post-course information:

You can attend one of our related courses or use the Additional Material links to further develop your R skills.

One-hour of follow-up time per course is available with the respective tutor if you would like any further guidance or discussion after a course. Please contact IT Learning Centre to arrange.

Housekeeping



- Exit the building if the fire alarm sounds
- Water cooler in the room and hot drinks available in registration area
- Toilets are located along the corridor
- The seats/monitors are adjustable
- If you have any concern or problem during the course please let me know



We are grateful to Dave Baker, Gabriele Pani and the administrative staff at IT Learning Centre for their support and assistance.

A series of related R courses are currently offered through the IT Learning Centre, developed and taught by:

Punam Amratia (Big Data Institute)

Rohan Arambepola (Big Data Institute)

Helen Lockstone (Wellcome Centre for Human Genetics)

Andre Python (Big Data Institute)

Ben Wright (Wellcome Centre for Human Genetics)

Similar courses were previously taught by Samantha Curle, who recently moved on from Oxford. We are grateful to all of them for discussions, contributions and ideas.

Related R Courses



- R: Kick-off
- R: Introduction to Basic Features
- R: Data Handling
- R: Visualisation
- R: Data Analysis courses coming soon

Getting Started with R



iT Centre
Learning

iT
services



Introductory Remarks



Understanding how to interact properly and carefully with R is vital to analyse your data correctly - mistakes can all too easily arise from oversights such as extracting the wrong portion of data or failing to spot data entry errors or inconsistencies before performing the analysis.

These can also be hard to detect in the output so a key habit to adopt is frequent checks of data objects and their contents, and making plots that will help spot issues or confirm if an output makes sense.

We will spend today's session introducing basic features of R and gaining some familiarity with using it as a programming language and to perform typical tasks associated with analysing data. Learning R is a long process but we hope this course will help you get started.

The RStudio interface



An interactive and easy-to-use interface with many features that make working with R easier:

(<https://www.rstudio.com>)

The screenshot displays the RStudio interface with the following components:

- Script Editor:** Contains R code for generating random data and plotting it.

```
1 # Write your script here
2
3 x = rnorm(100, 1, 1)
4 y = rnorm(100, 1, 1)
5
6 p = plot(x, y)
7
8 print(p)
```
- Environment Pane:** Shows the current environment with variables `x` and `y` of type `num` (numeric) and length 100.
- Console:** Displays the output of the script execution, including a workspace load message and the execution of `source()` commands.
- Plots Pane:** Shows a scatter plot of `y` versus `x`, with both axes ranging from -1 to 4.

The RStudio interface



Console for entering
R commands

A screenshot of the RStudio interface. The main window is titled 'my_script.R' and contains the following R code:

```
1 # Write your script here
2
3 x = rnorm(100, 1, 1)
4 y = rnorm(100, 1, 1)
5
6 p = plot(x, y)
7
8 print(p)
```

The right-hand side of the interface shows the 'Environment' pane with the following data:

Variable	Value
p	NULL (empty)
x	num [1:100] 2.46 0.287 -0.193 1.568 0.914 ...
y	num [1:100] 0.543 0.696 2.439 2.238 3 ...

Below the environment pane is a plot window showing a scatter plot of x vs y. The x-axis ranges from -1 to 4, and the y-axis ranges from -1 to 3. The plot contains approximately 100 data points, which are scattered across the plot area. The console window at the bottom shows the following output:

```
Natural language support but running in an English locale
R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.
Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.
[Workspace loaded from ~/.RData]
>
>
> source('~/.active-rstudio-document')
NULL
> source('~/.active-rstudio-document')
NULL
>
```

The RStudio interface



Script, data
tables etc



The screenshot displays the RStudio interface with the following components:

- Script Editor:** Contains the following R code:

```
1 # Write your script here
2
3 x = rnorm(100, 1, 1)
4 y = rnorm(100, 1, 1)
5
6 p = plot(x, y)
7
8 print(p)
```
- Environment Pane:** Shows the Global Environment with the following values:

Variable	Class	Attributes
p	NULL	(empty)
x	num	[1:100] 2.46 0.287 -0.193 1.568 0.914 ...
y	num	[1:100] 0.543 0.696 2.439 2.238 3 ...
- Console:** Shows the R startup message and the execution of the script:

```
Natural language support but running in an English locale

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

[Workspace loaded from ~/.RData]
>
>
> source('~/.active-rstudio-document')
NULL
> source('~/.active-rstudio-document')
NULL
>
```
- Plots Pane:** Displays a scatter plot of the generated data with x and y axes ranging from -1 to 4.

The RStudio interface



The screenshot displays the RStudio interface with the following components:

- Script Editor:** Contains R code for generating random data and plotting it.
- Console:** Shows the execution output, including R help text and the result of the `source()` function.
- Environment Pane:** A yellow-bordered window showing the current workspace. It lists objects `p`, `x`, and `y` with their respective data types and values.
- Viewer Pane:** Displays a scatter plot of the generated data with axes labeled `x` and `y`.

Clickable list of objects in memory



The RStudio interface



The screenshot displays the RStudio interface with the following components:

- Source Editor:** Contains an R script named `my_script.R` with the following code:

```
1 # Write your script here
2
3 x = rnorm(100, 1, 1)
4 y = rnorm(100, 1, 1)
5
6 p = plot(x, y)
7
8 print(p)
```
- Environment Pane:** Shows the Global Environment with the following values:

Variable	Value
p	NULL (empty)
x	num [1:100] 2.46 0.287 -0.193 1.568 0.914 ...
y	num [1:100] 0.543 0.696 2.439 2.238 3 ...
- Console:** Shows the R startup message and the execution of the script:

```
8.8 (Top Level) >
Natural language support but running in an English locale

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

[Workspace loaded from ~/.RData]
>
>
> source('~/.active-rstudio-document')
NULL
> source('~/.active-rstudio-document')
NULL
>
```
- Plots Pane:** Displays a scatter plot of `x` vs `y` with axes ranging from -1 to 4. The plot is highlighted with a purple border.

← Plots, files, packages, help etc...

RStudio on Windows



The screenshot displays the RStudio Windows interface with the following components:

- Script Editor:** Contains R code for creating vectors and plotting a sine wave.
- Console:** Shows the execution of the code, including the output of a plot and the help text for the `plot` function.
- Environment Pane:** Lists variables in the Global Environment, such as `bmi`, `v`, `x`, `x2`, `y`, and `z`.
- Help Pane:** Displays the documentation for the `plot` function, including its description, usage, and arguments.

```
1 x <- 5:20
2
3 1+2
4
5 x <- c(0,2,-1,pi,10)
6 x
7 x <- c(0:5,10:15)
8 x
9
10 y <- c(1, 2, 3, 4, 5)
11 x * y
12
```

```
> par(mfrow=c(1,2))
> plot(x,y)
> plot(x,y, las=1, xlab="x", ylab="sin(x)", type="l", main="Trigonometric functions")
>
> help.start()
If nothing happens, you should open
'http://127.0.0.1:21471/doc/html/index.html' yourself
> help("plot")
> ?(plot)
> help("plot")
> v <- -50:50
> v
 [1] -50 -49 -48 -47 -46 -45 -44 -43 -42 -41 -40 -39 -38 -37 -36 -35
[17] -34 -33 -32 -31 -30 -29 -28 -27 -26 -25 -24 -23 -22 -21 -20 -19
[33] -18 -17 -16 -15 -14 -13 -12 -11 -10 -9 -8 -7 -6 -5 -4 -3
[49] -2 -1 0 1 2 3 4 5 6 7 8 9 10 11 12 13
[65] 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29
[81] 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45
[97] 46 47 48 49 50
> x <- seq(-2*pi,2*pi,length=101)
> y <- sin(x)
> plot(x,y)
> plot(x,y, las=1, xlab="x", ylab="sin(x)", type="l", main="Trigonometric functions")
> help("plot")
>
>
>
>
>
>
>
>
>
>
```

Environment

Variable	Value
bmi	24.5
v	int [1:101] -50 -49 -48 -47 -46 -45 -44 -43 -42 -41 ...
x	num [1:101] -6.28 -6.16 -6.03 -5.91 -5.78 ...
x2	num [1:12] 0 1 4 9 16 25 100 121 144 169 ...
y	num [1:101] 2.45e-16 1.25e-01 2.49e-01 3.68e-01 4.82e-01 ...
z	0.333333333333333

Generic X-Y Plotting

Description

Generic function for plotting of R objects. For more details about the graphical parameter arguments, see [par](#).

For simple scatter plots, `plot.default` will be used. However, there are `plot` methods for many R objects, including [function](#)s, [data.frames](#), [density](#) objects, etc. Use methods (`plot`) and the documentation for these.

Usage

```
plot(x, y, ...)
```

Arguments

- `x` the coordinates of points in the plot. Alternatively, a single plotting structure, function or any R object with a `plot` method can be provided.
- `y` the y coordinates of points in the plot, optional if `x` is an appropriate structure.
- ... Arguments to be passed to methods, such as [graphical parameters](#) (see [par](#)). Many methods will accept the following arguments:

`type` what type of plot should be drawn. Possible types are

- "p" for points,
- "l" for lines,
- "b" for both,

R basics: the R environment



R is a very interactive environment – commands can be entered into the console one by one, and each is interpreted and executed by R in real time. R is a high-level programming language, meaning that it is fairly human-readable.

Depending on the command, different things may happen – a new object might be created, some output displayed, computations performed, plots generated etc.

If a command is not valid in the way it is constructed (its syntax), R will print an error message to the screen. These can sometimes be hard to interpret but particularly common culprits are simple typing mistakes or quotes and brackets, whether they are in the wrong place, missing, or not in pairs (e.g. missing a closing bracket).

We will start by entering a few simple commands and discuss what is happening. For now we will work directly in the console but later we will save our code in a script.

R basics: notation



Throughout these slides, text in red denotes commands that you can type in your R session (it is good to type them directly rather than copy and paste to get used to syntax of commands and expressions, at least to begin with).

Note that commands are shown including the prompt sign `>` but this appears by default in the R console and you don't need to type it yourself. For example

```
> x <- 5
```

should be entered at the command line as:

```
x <- 5
```

An extra `>` as well as the in-built R prompt will give an error:

```
> > x <- 5
```

```
Error: unexpected '>' in ">"
```

Commands are shown in red, and output in blue (this is the convention in the standard Windows R GUI but differs in other interfaces)

R basics: command line



Getting used to the command line (in red):

Prompt [Whitespace] Command [press Return key]

```
> 1+2
```

```
[1] 3
```

```
>
```

The output of this command (shown in blue) is printed directly to the console before the command prompt appears again. Elements of the output are indexed by the square brackets (in this case there is only one but sometimes longer lists of elements are returned).

Some commands do not produce any output; after executing the line of code, the prompt immediately appears again - it may not look as though anything has happened but it is likely something has changed. It is important to know what each command has done and it was as intended. The new prompt tells you that R is ready for you to enter another command.


R basics: data structures



To do anything useful in R, we need to use **objects** to hold data or information and perform various operations on them. The terms ‘object’, ‘variable’ and ‘data structure’ can all refer generally to objects created in R.

Although *variable* is a widely used programming term and would be the preferred term in certain situations, I will use *object* as a general term throughout to refer to any of R’s data structures. These include vectors, factors, matrices, dataframes and lists. We’ll focus on vectors initially and meet dataframes later on.

R basics: assignment



The following command assigns the value 1 to a new object we create and name 'x'.

```
> x <- 1
```

The assignment operator is <- and running this command creates a new object in R's memory

Inspect the contents of the new object

```
> x
```

```
[1] 1
```

Note that the contents of an object will be overwritten if later assigned something else:

```
> x <- 4
```

```
> x
```

```
[1] 4
```

We can also perform operations directly on the object (note the object itself does not change):

```
> x * 2
```

```
[1] 8
```

Unless we were to re-assign the output to it....

```
> x <- x * 2
```

R basics: naming objects



R is case sensitive so x and X are different:

```
> X  
Error: object "X" not found
```

Object names are chosen by the programmer – informative names are helpful for several reasons. You can use capitalization, `_` or `.` to separate parts of an object name but they cannot contain spaces, nor start with a number. To avoid confusion or potential issues, it is also best not to give them the same name as an R function, which have their own defined names. An object named `raw_data`, `raw.data`, `rawData` (or even `d.raw` for minimal typing!) is fine, but trying to assign a value to a variable named 'raw data' will give an error because R cannot parse it correctly:

```
Error: unexpected symbol in "raw data"
```

Elsewhere R ignores whitespace so the following commands are equivalent:

```
> x<-4+3  
> x <- 4 + 3
```

R basics: command syntax



```
> x <- c(1, 2, 3, 4, 5) # this creates a vector named 'x' containing some numeric values
```

If we forget the closing bracket before pressing enter, a + sign indicates the command is incomplete:

```
> x <- c(1, 2, 3, 4, 5  
+ )  
>
```

- If we can't simply continue our command, use Esc or control-C to return to the prompt and start again
- There is also a useful command recall option – you can use the up/down arrows to scroll through previously entered commands, which can be edited or re-run to save typing again
- Rstudio and some text editor programs highlight different parts of the syntax in different colours and automatically close brackets and quotation marks to help eliminate typing mistakes

Data Types and Structures in R



R basics: objects



Objects can be created in many different ways and hold different kinds of information. Unlike other programming languages, there is no need to initialise a variable or object in R (define it before first use) – it can simply be created and used directly. R also automatically decides which of its data structures and types are most appropriate for the data given, rather than being explicitly specified by the programmer.

We'll work through some examples and look at ways to access or manipulate the data contained within an object. Be aware that the type (class) of an object and data it contains (numeric, character etc) can affect how it is treated by R.

R basics: creating vectors



There are many shortcuts in R to avoid tedious or error-prone steps. When we created our small example vector containing the numbers 1 to 5, we issued the command:

```
>x <- c(1,2,3,4,5) # this tells R to concatenate these 5 numbers
```

Equivalently we could write:

```
> x <- 1:5
```

This is very handy if we wanted a much longer vector such as 1 to 100, or 1 to 1000,000

We can also put together non-consecutive strings of numbers or a mixture:

```
> x2 <- c(1,3,5,7,9)
```

```
> x3 <- c(1:5, 7, 9, 10:15)
```

If we need to create a sequence of numbers, the function 'seq' is very useful.

```
> seq1 <- seq(from=1, to=99, by=2)
```

```
> seq2 <- seq(from=0, to=1, by=0.01)
```

R basics: vectors



Vectors are one-dimensional objects; in the case of the object we created 'x', it has length 5.

There is an in-built R function called 'length' that we can use to check how long any given vector object is:

1	2	3	4	5
---	---	---	---	---

```
> length(x)
[1] 5
```

If we change what is assigned to x, the length of the vector is automatically adjusted:

```
> x <- 1:10
> length(x)
[1] 10
```

1	2	3	4	5	6	7	8	9	10
---	---	---	---	---	---	---	---	---	----

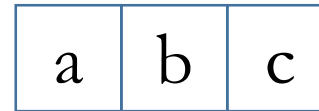
Vectors are R's primary object type and many computations are highly efficient because they operate on the whole vector at once, rather than element by element.

R basics: vectors



Vectors can contain numeric or character data (or both). We can create a new vector, `y`, containing the letters 'abc':

```
> y <- c("a", "b", "c")
```



In Rstudio's top-right panel, we see details of all the objects that have been created in the current session and are available to use. Note the differences between `x` and `y`.

We can also see how R has automatically treated them differently by checking the *class* of the objects directly:

```
> class(x)
```

```
[1] "integer"
```

```
> class(y)
```

```
[1] "character"
```

Now try running the following command:

```
> y <- c(a, b, c)
```

What do you think R has tried to do and why does it result in an error message?

Accessing elements of an object



Square brackets are used to refer to specific elements or subsets of a vector, factor, matrix or dataframe. Creating a new vector as an example:

```
> x <- c(1:5, 10:14)
```

1	2	3	4	5	10	11	12	13	14
---	---	---	---	---	----	----	----	----	----

```
## extract 3rd element
```

```
> x[3]
```

```
[1] 3
```

```
## extract alternate elements
```

```
> x[c(1, 3, 5, 7, 9)]
```

```
[1] 1 3 5 11 13
```

```
## extract subset of elements
```

```
> x[3:6]
```

```
[1] 3 4 5 10
```

R basics: object classes



R will decide the most appropriate way to store the data it is provided with, and there are ways to convert between different object structures and classes if needed. To give more examples of how data is interpreted by R, run the following and note the results (discuss with a neighbour):

```
> x2<-c(1:5, 6.5)
> class(x2)
> x3 <- c(1:5, 6.5, "a", "b", "c")
> class(x3)
```

This gives some idea of R's internal rules. Because the way data is being handled by R is important for both performing computations correctly and the source of many error messages, it is useful to be familiar with the common data types. Some functions, such as computing a mean for example, require numeric data objects to operate on:

```
> mean(x2)
> mean(x3)
```

R basics: object classes



The RStudio panel that shows existing objects and displays information on their contents is invaluable. Not only does it save writing separate commands to check these details, it can help you check:

- that your object has been created correctly and contains what you wanted it to
- how R will treat the object internally (when functions are applied to it)
- possible reasons for an error message
- spot any changes that happen to your object (intentionally or otherwise)
- that you do not have too many and/or poorly named objects that could lead to mistakes

TIP Sometimes you are testing things out and creating lots of objects – that’s fine but it’s good to start a new session when running or checking your final code to be sure previous objects do not affect it in any way. Sessions can also be cleaned up by deleting objects with the command:
`rm(object_name)`

R basics: functions



There are hundreds, probably thousands, of in-built functions in R. Some you will use very often and others rarely or never. There are always several ways to do the same thing in R, using closely-related functions.

Examples of the functions we have used so far include 'length', 'mean', 'class'. The name of the function is followed by a pair of braces (), within which specific information can be provided to the function. These are known as *arguments*, and enable the function to be used in a flexible way. In the case of length, the argument supplied is the name of the object we wish to find the length of. The 'length' function is only applicable to vectors (or factors) and does not work on other data types such as matrices or dataframes. We can find this information and details of how to use a function via the relevant help page (these load in the lower right panel of Rstudio for convenience).

```
> help(length)
```

We can check the length of another vector object simply by changing the argument:

```
>length(seq1)
```

```
[1] 50
```

R basics: the working directory



There are many useful functions in R for manipulating data stored in vectors, matrices or dataframes - best introduced through practical exercises we will do after the break.

One important concept before we get started is the **working directory**. If we want to read in data from existing files or create new ones to save any plots or analysis results, R needs to know where to find/save them.

```
> getwd() # tells us the current working directory
```

```
> setwd() # allows us to set a different working directory
```

For the latter, a path to the directory needs to be provided in quotes within the brackets. The path can be full or relative:

```
> setwd("C:/Users/jbloggs/data")
```

```
> setwd("./data") # assuming the current working directory is  
C:/Users/jbloggs (a level above the 'data' directory)
```

R basics: the working directory



The working directory can also be set and changed via the menu bar:

- Using the R console you find 'Change dir...' under the File menu
- In RStudio, the Session menu has an option 'Set working directory'.

From these you can navigate to the desired directory and select OK to set it as the working directory. Note that this happens behind the scenes and nothing will appear to happen on your screen - you can use `getwd()` to confirm if you have set it as intended.